DOCUMENT RESUME

ED 076 702 TM 002 711

AUTHOR	Ebel, Robert L.
TITLE	The Future of Measurements of Abilities II.
PUB DATE	27 Feb 73
NOTE	22p.; Speech given before the annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests; Aptitude Tests; *Educational Testing; Intelligence Tests; *Measurement Goals; *Measurement Instruments; Speeches; *Student Evaluation; Test Interpretation; Test Validity
IDENTIFIERS	Thorndike (E L)

ABSTRACT

The views of E. L. Thorndike on the future of measurements of abilities, expressed 25 years ago, are summarized, and the future of measurements of abilities as it appears now is examined. Opportunities for improvement now arise mainly from increasing social concern for effective education. Measurement technology has developed rapidly and cannot continue at the same pace. Newer instructional technologies will not be the most widely used because they are costly, impersonal, inflexible, and less learner-oriented. Formative evaluation can supplement but not replace summative evaluation, and criterion-referenced testing can supplement but not replace norm-referenced testing. The concept of mastery learning cannot be applied rigorously to most tests of abilities; the learning of any complex skill of understanding is always incomplete. Social concern is evidenced in public demand for accountability and governmental desire to allocate funds more equitably. A serious problem in the use of tests of ability is what to measure--what the proper roles are for intelligence tests, tests of general mental abilities, critical thinking tests, tests of creativity, and tests for affective outcomes. Some problems in determining test validity are created by asking the wrong questions and by not recognizing that each different test measures a somewhat different ability. The term "construct validity" is used loosely and with a variety of meanings. Employment tests have been questioned by courts recently, and their validity must be proved. Another major problem is that of developing the necessary tests in quantity and at low cost. (KM)

FILMED FROM BEST AVAILABLE COPY

ED 076702

The Future of Measurements of Abilities II[1]

Robert L. Ebel

Michigan State University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

1. Introduction

The title of this paper was borrowed from one published exactly twenty-five years ago, February 1948, in the British Journal of Educational Psychology.[2] The author was an American educational psychologist, Prof. Edward L. Thorndike, then near the close of a long, productive, distinguished career. Most of you know that Robert L. Thorndike, the son of Edward and like him a distinguished educational research worker, is president-elect of this association.

All of us here owe a great debt to formative leadership of E.L. Thorndike in educational research and educational measurement. That would be reason enough for us to honor his memory on this occasion. But there is another, even more persuasive reason. In this decade of the seventies measurement not only has unprecedented opportunities to serve the cause of better education. It also is confronted with some very serious challenges to its methodological soundness and to its social utility. If educational measurements are to make optimum contributions to the progress of education in America we need all of the wisdom we can muster regarding both its possibilities and its limitations. That is why I propose that tonight we take a second look at the future of measurements of abilities, twenty five years later.

---

[1] Presidential address, American Educational Research Association, New Orleans Louisiana, February 27, 1973.

[2] Volume 18, pages 21-25.

2. Professor Thorndike's views

Let me begin by reviewing briefly what Professor Thorndike had to say. One might guess from the title that he would make some predictions of what was likely to happen. Not so. He was more concerned with what ought to happen, with what wisdom and hard work could make to happen. His article was essentially a set of prescriptions or suggestions for the production of better tests of ability.

He began by naming three qualities that, he said, all competent students would agree any measurements of abilities ought to possess as much of as possible. Are there any competent students of the measurement of abilities in this room? Then of course you know what he was talking about. You know that he did not name validity, or reliability, or norms, or convenience in use. You know that, instead, the three crucial qualities he named were objectivity, adequacy and purity. By objectivity he meant the closeness of agreement in the scores given to the same examinee by different examiners. By adequacy he meant how nearly the test measured all of the ability in question. By purity he meant how little the test measured of anything other than the ability in question.

How important do these three qualities seem to you? Would you regard them as essential in any good test of ability? Taken together, they seem to provide a fairly good basis for what we call content validity. The concepts of adequacy and purity seem to me to be useful. While I am not prepared to let them displace validity or reliability in my vocabulary, I am ready to let them warn me to look carefully at the adequacy and the purity of any tests I use.

Prof. Thorndike did not regard such things as ability in French, in chemistry, in music, or in athletics, or ability with ideas or with mechanisms, as unitary faculties or essences. He regarded measurements of them as essentially inventories. Hence he suggested this sequence of steps in test development.

First, prepare as adequate (and pure) an inventory of the ability in question as

a concensus of experts can develop.

Second, devise a criterion test that will correlate .95 or higher with the best weighted score from the total inventory.

Third, devise multiple working tests that will correlate .95 or higher with the criterion test.

While I share the belief that human abilities are seldom, if ever, unitary faculties or essences, I have some difficulty with the notion that the components of most abilities are discrete and finite enough to allow an inventory of them to be made. This is a point we will return to later. Those correlation coefficients of ".95 or higher" worry me a bit too. I have seldom encountered a test of any mental ability that could produce a correlation of .95 with anything. Finally, the distinction Prof. Thorndike made between criterion test and working tests is intriguing, but it too raises questions. I suspect that it may have been borrowed from physical measurement. There is, or was, a carefully guarded platinum - iriduim bar in Paris that defined the international standard of length, the meter. In tens of thousands of shops, laboratories and classrooms around the world there are hundreds of thousands of more or less battered replicas of that standard in use as working tools of measurement. It is tempting to imagine that this nice combination of protected precision and unlimited replication might be copied in mental measurement. But where, in the realm of human abilities is there anything comparable to length in simplicity of conception or operational definition? What hope is there of getting educators to agree to use a single definition of any practically useful human ability such as ability to read, write or calculate? In any event, not much use of this suggestion has been made in the quarter century since it was offered.

Professor Thorndike made several other points in his article which merit at least brief comment here. He said that of several equally valid tests, those of lowest reliabilities are most promising, because their validity can be increased most by

combining or lengthening them. This is surely true in principle. In practice it is most uncommon to find tests of demonstrably equal validities that differ appreciably in reliability.

He said that attempts to measure such things as "imagination" and "leadership" are frustrated by our ignorance of what we are trying to measure. I would agree, and add that words we use to describe behavior sometimes are transformed, subtlely and with little justification, into names for causes of that behavior. What we often try to measure, when we deal with things like "imagination" and "leadership" is not the extent of the manifest behavior, but the presumed capability or propensity for it. If, as often seems to be the case, the manifest behavior has manifold causes, many of which lie not in the behaver but in the situation in which he finds himself, it is not surprising that attempts to discover the cause in him often flounder.

Thorndike claimed that a purer (i.e. more meaningful) measure of ability in a foreign language could be obtained from a test based on phrases and short sentences than from one based on a long passage of connected discourse. In the latter type of test, he argued, general intelligence, general background, or general reading ability may be a substantial score-influencing factor that has little to do with foreign language ability. I agree. There are situations, as Thorndike noted, where contamination of measures of a specific ability with general intelligence is advantageous, but we ought always to remember that the contamination exists.

Finally, Thorndike in 1948 took note of the investigations of Hotelling and Thurstone into factor analysis. He concluded that this work "... has not so far increased our equipment of adequate tests of pure abilities much if at all." Then he added, "I do not require them (that is, tests of pure abilities) of the future partly because I do not believe the mind is composed of such, and partly because in any case there are more urgent needs." The more urgent needs, I take it, are for

tests of practically useful abilities; abilities that contribute to success in life or success in some aspect of the world's work; abilities that good teachers try to teach and good students to learn. My impression is that at least some of us in this room, twenty-five years later, would support his judgment.

3. Opportunities and problems today

Let us turn our attention now from Thorndikes observations and recommendations to look at some of the opportunities and problems that face us today. Like him we probably should be less concerned with forecasting what is going to happen than in determining what ought to happen and how we can make it happen. I have some ideas on this matter to present for your consideration. You may or may not find them reasonable and acceptable. If our opinions do differ, let us not therefore think the less of each other. Recall what John Milton said in his classic defense of freedom of speech.

> "Where there is much desire to learn there of necessity will be much arguing, much writing, many opinions; for opinion in good men is but knowledge in the making."[3]

The first conclusion I have reached and now hold, at least tentatively, is this. The opportunities that this decade presents to us for improvements in the measurement of abilities arise mainly from increasing social concern for effective education, and not, I believe, from prospect for the success of radical innovations in the theory or practice of educational measurements. The first three quarters of this century have witnessed fantastic developments in measurement technology. For this progress we are indebted to men like E.L. Thorndike, T.L. Kelley, Alfred Binet, Charles Spearman, Arthur Otis, Ben Wood, L.L. Thurstone, Ralph Tyler, E.F. Lindquist, John Flanagan, Harold Gulliksen and Fred Lord, among others.

---

[3] Milton, John Aeropagitica, 1644.

It would be unrealistic to expect the kind of progress they made to continue indefinitely. Further, the problems which are holding back the more effective utilization of measurement in education seem to me not to be primarily theoretical or technological problems. Nor do any recent innovations in those fields seem to promise effective solution of the most serious measurement problems.

To be specific, I do not foresee extensive use of tests built by computer from item banks, nor of individualized, variable sequence testing controlled by computers. I see some theoretical value, but little practical utility in systematic item development via linguistic analysis of sentences used in instruction. The use of document readers to extend the range of item types that can be scored by machine likewise seems to me unlikely to have much impact on the practice of educational measurement.

The development of instructional technology, of systematically programmed instruction, of individually prescribed instruction, of computer assisted instruction, of sequential mastery learning, has led to emphasis on formative evaluation as well as, or in place of summative evaluation; on criterion referenced rather than on norm referenced tests; on learning to absolute mastery rather than on the partial learning so prevalent in conventional systems of education. No doubt developments along these lines will continue and ought to be encouraged. In some situations they may prove to be substantially more effective than previous techniques have been. However there are several reasons why the newer instructional technologies seem to me unlikely to become the prevalent modes of instruction in the foreseeable future.

In the first place, development of high quality materials for systematic instructional strategies tends to be costly. Second, students often find impersonal systematic instruction somewhat dull, after the novelty has worn off. Rigidly prescribed behavior becomes irksome. Third, the necessarily detailed preplanning of each step in the instructional process prevents the kind of flexibility available

in conventional instruction, flexibility that allows unexpected problems to be solved, unexpected opportunities to be seized. Finally, the emphasis in systematic instruction is on the instruction, not on the learner. The more obediently he follows directions the better. As compared with conventional instruction, systematically pre-planned instruction calls for less initiative, less inquiry, less self direction, less self evaluation, on the part of the learner. If learning does not occur or is inadequate, the blame is likely to go to the instructional system, not to the learner.

It is interesting to note that the free-school movement has moved in precisely the opposite direction in its attempt to improve on conventional schooling. Instead of specifying the steps in the instructional process in great detail, as systems of systematic instruction try to do, the free school leaves the direction of this process very largely in the hands of the learner himself. In the free school there is a maximum of freedom and flexibility. There is a minimum of carefully planned steps to be taken in pursuit of carefully specified goals.

Perhaps a happy medium, an optimal instructional strategy, exists somewhere between these two extremes of total control and absolute freedom. Perhaps finding that happy medium is one of the essential requirements for effective teaching. Perhaps learning proceeds best when two requirements are met:

1) The learner brings to the task a strong desire to learn, a willingness to do the hard thinking that learning often requires, and
2) The school provides a good learning environment which consists of (a) capable enthusiastic teachers and (b) a generous supply of good books and other aids to learning, (c) opportunities to interact with other good students, and (d) a system for recognizing and rewarding achievement in learning.

The school is responsible for the instruction. The student is responsible for the learning. Neither can do the whole job alone. Perhaps systematic instruction fails to stress sufficiently the student's responsibility for learning. Perhaps the free school movement fails to stress sufficiently the school's responsibility for instruction.

4. Some implications of new instructional technologies

Let us turn our attention now to some of the implications for measurement of the newer instructional technologies. In particular let us consider formative evaluation, criterion referenced testing and mastery learning.

The use of formative evaluation to help in the guidance and facilitation of learning is surely to be encouraged. In ordinary situations it does not replace summative evaluation, that is, the assessment of achievement in learning. Further, in many situations, formative evaluation can be handled quite casually and informally. Indeed, while formative evaluation may be a recently popular term, it refers to an ancient and honorable activity of good teachers everywhere: the observation of student progress and difficulty in learning, followed by adjustment in instructional procedures to improve that progress.

Just as formative evaluation can supplement but not replace summative evaluation, so criterion referenced testing can supplement but not replace norm referenced testing. Tests of ability can, and often should, yield two different kinds of scores that convey two different kinds of meaning. The first is content or criterion - related meaning. How much of this area of knowledge does the examinee command? How much of that ability does he possess? The second is relative, normative meaning. With respect to this area of knowledge or that ability where does the examinee stand in relation to the scores made by a specified group of his peers?

Some measurement specialists disagree with the conception of a criterion - referenced test score presented above.[4] They contend that the purpose of a criterion referenced

---

[4] Block, James H. "Criterion-referenced Measurements: Potential" School Review, 79(1971) 289-298

test is to reveal what the examinee knows, not how much he knows. Instead of reporting the number of items correctly answered, they would simply list the items answered correctly. In this conception a criterion - referenced test is more of an inventory than of a measuring device. The inventory can be useful in remedial instruction, but it does not provide a measurement of the overall success of a pupil's efforts to learn.

There is another difficulty with this method of assessing achievement. It places full faith and gives full credit to a pupil's answer to a single test item. Now as most test specialists know, the score on a single test item is likely to be quite unreliable. To make decisions on what particular things a student knows or does not know on the basis of a single response from the student is only a little better than to do so by tossing a coin. Thus when a criterion referenced test is used to determine exactly what a student knows or does not know, the determination is likely to be often in error.

On the other hand, if a criterion referenced test is used to determine how much a student knows, the problem of units of measurement arises. In some areas of learning, such as the basic facts of addition and multiplication, or the spelling of words in a particular list, achievement does come in fairly discrete, non-overlapping packages, and the area of learning has finite limits. In most areas of learning, however, this is not the case. The knowledge is a complex structure of concepts and relationships. There are no discrete units of knowledge. The number of true propositions that can be used to express that knowledge is almost infinite. Seldom can anyone assert truly that he knows all there is to know about the subject.

In these circumstances it is very difficult to determine exactly and objectively how much of an area of knowledge any examinee can command, or how much of an ability he possesses. You can report how many of the items on a particular test a particular examinee answered correctly, and can express this number as a percent of the total

number of items in the test. But ordinarily there is no sound logical reason to regard the percent of items correctly answered as a reasonable estimate of the percent of mastery the student has achieved. Indeed it is almost impossible in most situations to attach any operational meaning to the phrase "percent of mastery."

For the same reasons, the concept of mastery can not be applied rigorously to most tests of abilities. The units are too indistinct and the total area to be dealt with is too indefinitely defined. Those who are said to have mastered an area have usually only attained some arbitrarily defined, imperfect grasp of it. The learning of any complex skill or understanding is always incomplete. Total mastery does not exist, and no "mastery model" for teaching can produce it.

5. Social concern for effective education

So much for our skepticism as to the likelihood of radical changes in the technology of testing in response to radical innovations in teaching strategies. We said earlier that the opportunities this decade presents for improvements in the measurement of abilities arise mainly from increasing social concern for effective education. Let us be more specific.

Things have not gone well in the public schools in recent years. While costs have escalated, the quality of the product seems to have deteriorated. Semi literate students have been graduated from high school. Disaffected students have disrupted classes and vandalized school buildings. Teachers have struck for higher pay and better working conditions. So the public has begun to ask for evidence that the money it is spending on education is being well spent. It has begun to demand that school boards and their agents, the school administrators and teachers be held accountable for the beneficial results of public investments in education. In response to their demands, state legislatures have turned to mandatory testing programs to provide some of the needed evidence. But the tests that are readily available or

can be produced easily often seem unequal to the task that is required of them. This gives us one strong incentive to seek improvements in measurement of abilities.

Another opportunity is provided by the need to evaluate the effectiveness of new instructional programs. A wide variety of new educational procedures have been initiated nationally and locally to solve a host of educational problems: to overcome early educational deficits; to prevent dropouts; to provide vocational training; to encourage college attendance by minority group students; and to improve instruction in various subject areas. The use of tests of abilities to evaluate educational programs is by no means new, but modern demands for program evaluation call for the development of more comprehensive and detailed evaluation techniques.

A third opportunity grows out of the desire of governments, federal and state, to allocate educational resources more equitably. This calls for more precise assessments of educational needs than have been made in the past. Again, tests of ability are being asked to provide data for these assessments.

Thus the opportunities for effective use of tests of ability abound on the contemporary educational scene. Those who dislike and mistrust tests, who are more conscious of their limitations than of their contributions are not happy with calls for more testing. The weakness of their position is that they seem to have nothing better than tests to offer. The need for evidence can hardly be denied. No more promising source of that evidence than is provided by tests of ability seems to be available. But tests do have limitations. Problems are associated with their use. All those who use tests need to be aware of these problems and of some solutions for them.

One of the most basic and serious problems which affect the effective use of tests of ability in education is the problem of what to measure. In particular what is the proper role of intelligence tests, of tests of general mental abilities, of critical thinking tests, of tests of creativity, of tests for affective outcomes?

Let us consider each of these in turn, starting with intelligence tests.

6. The role of tests of intelligence and of general mental abilities.

A case can be made for giving up the notion that any of the currently available individual or group tests of intelligence are measures of native intelligence. The tasks which compose them invariably require the examinee to do something that he has learned to do, not something he was born knowing how to do. While intellectual ability must have a biological basis, and while it is reasonable to assume that individuals may differ in the quality of that basis, there is as yet no answer to the question, "What biological difference makes normal people differ in intelligence?" When we set out to measure native intelligence we have no idea what it is that we are trying to measure, and no way of knowing whether or not, or how well, we have succeeded in measuring it.

Now there is nothing at all wrong, indeed there is very much that is right, with the use of intelligence tests as measures of general ability to learn. But there is very much that is wrong, educationally with the assumption that a child who scores low on an intelligence test is biologically limited in learning ability. There is very much wrong socially with the assumption that a cultural group which scores lower on such tests than another is less well endowed biologically than the other. These are assumptions whose truth can not be tested in the present state of neurophysiological knowledge. They are not necessary to explain the results we get from intelligence tests. They can do, and have done much harm, educationally and socially. Until we know much more about native intelligence, its origin and nature, than we do today, we will be well advised not to claim that any of our current intelligence tests measure it.

Is it the major task of education to cultivate a person's general mental abilities and the major task of good tests to determine how well they have been developed? Are these abilities few in number, very general, and highly abstract like ability to

analyze, ability to synthesize, ability to evaluate, ability to measure, ability to apply, ability to experiment, or even ability to think? Or are they multitudinous in number, very specific and highly concrete, like ability to spell "extraterrestrial" or "Afghanistan", like ability to multiply two common fractions, like ability to read the menu in a French restaurant, like ability to differentiate between reliability and validity, and so on.

It seems clear to me that the abilities with which teachers and test makers need to be concerned are of the latter kind, concrete, specific and very numerous. I am persuaded that all useful learning begins with particular learnings, and that a general ability, like the physicians ability to diagnose a patients ailment consists entirely of a host of specific diagnostic abilities. It is a host of concepts and relationships, facts and generalizations, that enables one to think. The quality of a person's thinking, I believe, is determined far more by the extent and quality of his knowledge of these concepts, relationships, facts and generalizations, than it is by any special thinking skills (whatever they might be) that he has developed.

7. Critical thinking and creativity

If the quality of a person's thinking depends on the quality of his knowledge, it is also true that the quality of his knowledge depends on the quality of the thinking that developed it. Some students, particularly college students, seem most reluctant to tackle the harder part of learning, the part that demands critical thinking in order for them to build their own private structure of knowledge. The easier part is getting the information input to think about. They take copious notes in class. They highlight abundantly the pages they read. In a last minute review before the test they try to fix in mind all they have heard and seen. What they omit doing is struggling incessantly to answer three questions: What does it mean? How do they know? Why is it so? What they fail to do is to transform the information they have received into a structure of knowledge. And if the test scores of such

students are disappointing, as they well may be, they may transfer blame to the professor who, they charge expected them to "memorize a bunch of facts." In this they reveal their own misapprehension of the process and product of learning. The method they have failed to use is the method of critical thinking. In our teaching we should exemplify it. In our testing we should reward it.

Human minds like the ones you and I possess are the storehouses of vast amounts of knowledge that we have gained from study and experience, of understanding, of know-how. We know very little about how it is stored, or how it sometimes makes itself available when we want it. But we are conscious of its existence and convinced of its value. About mental processes, and the general mental abilities these processes might support, on the other hand, we know very little. Hence in teaching and testing it would seem to make sense for us to focus our attention on and to direct our efforts toward the cultivation, and the assessment, of useful knowledge and understanding. It would seem to make sense for us to have very little to do with hypothetical mental processes or general mental abilities.

Critical thinking, I believe, is not a kind of thinking. It is a use to which thought can be put. It is not something that students have to be taught <u>how</u> to do. It is something that they must be persuaded is worth doing. I do not believe that we need tests of critical thinking ability. What we do need is subject matter tests that call for understanding and application, for demonstration of command of knowledge. Only those who are used to thinking critically are likely to have solid and enduring structures of knowledge.

Next, what of the teaching and testing of creativity? To be creative of great works of art or literature, of scientific discoveries or technological inventions is admirable. The few who succeed in being notably creative, and whose success is noticed, richly deserve the honor accorded them. But I would insist that ordinary people like us are creative too in the ideas we have, the plans we make and carry out,

the problems we solve. Only prizes for creativity are seldom awarded us. Our creations are not perfect enough or important enough. It is not in being less creative that we differ from famous men and women. It is in being less excellent, or perhaps less lucky. Those who try to teach people to be creative in general, or test for creativity in general, seem to me to be chasing a will-o'-the-wisp. There is no good reason I know of to believe that those who excell in creative accomplishments owe their success to a super abundance of general creative ability or talent. Creative achievement seems always to depend on special abilities, on special opportunities, on special efforts.

8. Affective Outcomes

Finally, what of the affective outcomes of education? Teachers and testers are sometimes blamed for neglecting them, for concentrating on cognitive outcomes, because these are easier to teach and to test, even though affective outcomes may be more important. I agree that affect is important. How I feel is almost always more important to me than what I know. It seems obvious that teaching and learning, even when directed at cognitive outcomes, always has affective by-products. Good teaching will aim to make these by-products contribute to the happiness, adjustment and goodness of the learner. What I do not agree with is the implication that schools can and should commit a large fraction of their efforts to the fostering of a pupil's affective development. Is there any school or college that is currently channeling a major portion of its resources into a program of affective education? If such a program could be devised, what would it look like? How do you go about affective education? What goals do you seek to attain?

The authors of the Affective Domain Handbook of the Taxonomy of Educational Objectives[5] clearly recognized, and were distressed by, the primitive state of the

---

[5]Krathwohl, D.R. et al. Taxonomy of Educational Objectives: Affective Domain. New York: David McKay Company, Inc., 1964.

art of affective education. They acknowledge the very proper reluctance of teachers to grade pupils on the basis of their success or failure in attaining designated affective goals. They sought to alleviate this problem by suggesting better tools for evaluation. Indeed, the major purpose of the handbook was to improve affective education. Perhaps the nine years since publication of the handbook is too little time for its influence to be observed. But if there are currently in existence or at an advanced stage in planning strongly supported programs of affective education that give promise of more substantial attainment of affective goals, I have not run across them.

It is interesting to note that the taxonomy of the affective domain has very little to say about feelings or emotions, which is what the dictionary says "affective" means. Instead the taxonomy identifies the affective domain with interests, attitudes, appreciations, character and values. All of these seem to me to have fairly substantial cognitive bases. The major categories of the taxonomy: receiving, responding, valuing, organization and characterization; seem to be more descriptive of behavior than of feeling. These terms too seem to imply, and to depend for their development, on cognitive competence.

Perhaps I am not alone in my uncertainty concerning the nature of affective goals, the appropriateness and probable success of efforts to reach such goals by direct instruction, the efficacy of tests in measuring the success of such efforts. Perhaps the reason so few schools seem to be doing much about affective education is that there is nothing much that is sensible and effective to be done. In any case, I am not inclined to give high priority to the measurement of affective outcomes as we look to the future of the measurement of abilities.

9. Problems of test validity

Let us direct our attention next to some of the manifold problems of test validity. Does the test in fact measure what it is supposed to measure? Recall that Professor

E.L. Thorndike said a good test of ability should have as much objectivity, adequacy and purity as possible. Such a test would indeed measure what it is supposed to measure. It would, in other words, be a valid test. Recall that adequacy and purity were to be judged relative to an inventory of all the components of the ability in question, and that the inventory was to be developed by expert judges familiar with the ability in question.

A test developed by these procedures, and meeting these criteria would seem to me to have a reasonable claim to be regarded as a valid test. But I do have some questions. Are the abilities we speak of organic entities that exist to be discovered and explored, as an island or an archeological site exist? Or are they simply linguistic categories for related ideas and skills, having no single true definition but instead a variety of definitions to serve different purposes in different situations? And what of the components of these abilities? Do they too exist to be discovered, or can they be defined in infinitely various ways?

My own considered opinion is that neither the abilities nor their components exist to be discovered. If this is true there is no single common inventory against which the adequacy and purity of every test of that ability can be judged. Each different test implies a somewhat different inventory of the components of the ability. Each different test measures a somewhat different ability.

If this is the case our concern for the validity of such tests may be misplaced. That is, instead of asking "How objectively, adequately and purely does this test measure the ability in question?" we should be asking, "How clearly and how meaningfully has the test developer described what the test does measure?" "How clearly are the criteria for item content and item quality specified, and how faithfully were the specifications followed?"[6]

---

[6]Ebel, R.L. "Must All Tests Be Valid?" American Psychologist 16(1961): 640-47.

The model of test validation which involves the correlation of test scores with criterion measures is in many situations a misleading, unhelpful model. It requires criterion measures of unquestionable validity against which the validity of questionable test scores can be demonstrated. But in many situations criterion measures of unquestionable validity are not available. Test developers sometimes fall back on construct validation in these situations, often with results that seem satisfactory if they are not examined too critically. It will be worth our while at this point to take a brief look at construct validity.

10. Construct validity

Part of the problem with construct validity is that the term is used so loosely, and with such a variety of meaning. One source goes so far as to say that content, concurrent and predictive validity are all specialized aspects of construct validity.[7] Another suggests that construct validity is based on logical inferences from relevant evidence.[8] That would seem to just about cover all of the bases, not only of test validation but of scientific method as well. If scores from a new intelligence test correlate highly with scores from an established test, says one source, the new test possesses construct validity. I would have called that concurrent validity. Again the claim is made that an art aptitude test can be shown to have construct validity if artists make higher scores on it than non artists. Does an aptitude test need construct validity as well as predictive validity? If it does, would comparison of the scores of artists and non artists validate it as an aptitude test?

The basic procedure of construct validation involves two steps. First, hypothesize the relations that should exist between scores on the test to be validated and measures of certain other abilities or traits. Second, collect data to test

---

[7] Beggs, Donald L. and Lewis, Ernest L. Measurement and Evaluation in the Schools. Houghton-Mifflin Co., Boston. In press.

[8] Committee on Test Standards AERA-NCMUE, Technical Recommendations for Achievement Tests. National Education Association, Washington, D.C. 1955 pp.16-19.

the hypotheses. To the extent that the hypotheses are confirmed, the test is validated. Now if the hypotheses were exact, quantitative hypotheses, and if they were derived from a rational quantitative theory of human behavior, these procedures could indeed establish the validity of the test. The trouble with construct validation as it is ordinarily practiced is that the hypotheses to be tested are not exact quantitative hypotheses, and they are not derived from any quantitative theory of human behavior, because no such theory exists. In ordinary practice a spelling test is claimed to have construct validity as a test of spelling ability if sixth graders make higher scores on it than fourth graders. What bothers me about this inference is that sixth graders will do better than fourth graders on almost any test. One author suggested that a test of study skills can be shown to have construct validity if overachievers make higher scores on it than do underachievers. Again, I suspect that the overachievers may make higher scores than underachievers on almost any test. Which is the best way of determining whether a test of study skills actually measures study skills: by looking at the tasks that compose it or by comparing the scores of underachievers and overachievers on it?

Physical scientists define their quantitative constructs operationally. They almost never use correlation with a criterion to demonstrate the validity of those constructs or of their measurements of them. In the measurement of mental abilities, it seems to me, we ought to follow their example more often than we do. Our measurements of abilities all need to be operationally defined. They do not all need to be validated. And the proportion that can be shown, or need to be shown, to possess real construct validity is miniscule. Some of the problems of test validation that perplex us are problems we ourselves have created, unnecessarily.

11. The validity of employment tests

The validity of measurements of ability has come under particular scrutiny in courts hearing cases of alleged discrimination in employment. In general, the courts

have held that any company which uses tests in selecting employees must be prepared to show a substantial relation between performance on the test and success on the job. On the face of it this appears to be a most reasonable requirement. Indeed it is reasonable in that small minority of routine jobs where "success" is simple to define and easy to quantify. But in the vast majority of jobs, the evaluation of degree of success is a complex matter, one that can not be done with complete objectivity and assurance, one that can best be done by the deliberation of expert judges. Apparently the courts are willing to accept such judgments when applied to the behavior and personal characteristics of a person after he has been hired. Apparently they are not willing to accept the same kind of judgments if applied before the person has been hired. One is entitled to wonder if there is any sound logical or empirical basis for this differentiation.

Many companies have found it to be inordinately difficult to provide the kind of evidence of test validity the courts demand, and have been forced to discontinue use of the tests. Surely there are instances in which the tests were in fact inappropriate to the job requirements, and ought to have been dropped. But there are other instances in which sound rational grounds exist for believing that the tests do measure important qualifications for job success. Dropping such tests actually contributes to unfair employment practices. The notion that empirical evidence of test validity avoids the need for exercise of subjective judgment is fallacious. Success must be defined before it can be measured or predicted. Reason can be almost as powerful when applied to the problem of who is likely to succeed as when applied to the problem of who has succeeded.

12. Problems of test development

Thus far we have considered in detail two major problems which tend to limit the effective use of tests of ability. One is the problem of determining what to measure. Another is that of determining whether, in fact, our test actually measures

it. A third major problem is that of developing the necessary tests in quantity and at low cost. Thousands of men and women are writing test items for use in classrooms or by civil service agencies, or for research studies. Some of these item writers have little if any special aptitude for the task. Few of them have had systematic special training for the task. It is not surprising that the tests they produce are sometimes of low quality. While the quality of a finished test depends at least as much on the skill that goes in to the item writing as on the sophistication of the test analysis, the test developer has much easier access to a good supply of the latter than of the former. I believe this is a problem that ought to be solved, and that can be solved once we recognize its importance, and start working hard to solve it.

It goes without saying that not all measurement specialists are skilled item writers. Nor do they all agree on the merits of various item forms. Some deplore the dominance of the multiple choice item, contending, without supporting evidence or convincing rationale, that the range of cognitive abilities that can be tested by using multiple choice items is seriously limited. Many continue to council against the use of true false items, despite some pretty good rational arguments and empirical evidence that true false tests can do the same job that multiple-choice tests can do, about as well as multiple choice tests can do it, and sometimes much more conveniently. Some of them persist in referring to all choice type items as recognition tests, without regard for the complex thought processes involved in choosing the best answers to such items. If misconceptions like these limit the effective use of tests of ability, as I think they do, we who specialize in educational measurements can not blame anyone else. They are our problems, and their solution is our responsibility.

13. Conclusion

There are many other current problems in the measurement of abilities that would be interesting and useful for us to explore if time allowed. There is the problem

of the alleged bias of conventional middle-class tests against racial minorities. There is the problem of proper interpretation of test scores, and of the inadequacy of many professionally trained educators to make such interpretations. There is the problem of public attitudes toward testing, which seem to gravitate to the extremes of uncritical acceptance or disdainful rejection. There are the problems of needless duplication, of invasion of privacy, and of self fulfilling prophecies. But time does not allow more than passing reference to these problems.

It is interesting to speculate on how Professor E.L. Thorndike might react to these problems if he could be here tonight. I think he would be pleased with the progress that has been made during the last twenty five years in the measurement of abilities. But he would be more interested, I think, in the variety of problems that remain to be solved. I suspect that with his characteristic energy he would be anxious to begin work on them, and that with his characteristic clarity of thought he would shortly work out solutions to some of them. Let us try to follow his good example.